**Joseph Davis**

IST 652 Final Project

**Objective**

The objective of this final project was to look at movie data and figure out what distinct characteristics will help a film both make money and get positive reviews.

**Data**

The data looked at for this project comes from MovieLens, a movie review platform. The dataset was taken from Kaggle user Rounak Banik at: <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset> . This dataset included 7 csv files however not all were used for analysis. The csv files were credits, keywords, links\_small, links, movies\_metadata, ratings\_small, and ratings. Credits didn’t hold any meaningful data and the small versions of the keywords and ratings csv files weren’t relevant. The keywords csv file contained keywords about the movies however it wasn’t used in this analysis. The ratings csv file included individual user ratings, including user\_id, movie\_id, rating, and timestamp. Finally, the movies\_metadata file contained the bulk of the useful data. It included the movie title, genre(s), budget, original\_language, overview, popularity, production\_company, production\_country, release date, revenue, runtime, vote\_average, vote\_count, and a few other attributes. Preprocessing of this data was extensive. When looking at movies from a money-making perspective, all budgets and revenues that were below $10,000 were removed. Additionally, all films released before January 1st, 2000, were excluded. The budget column was a string originally with some jpg files about 4000 rows deep that needed to be removed. Addition ROI and net revenue columns were created. Further processing to answer individual questions was done and will be described later. In addition to these changes, irrelevant columns such as imdb\_id were removed leaving us with a final dataframe of 16 columns and 5853 rows.

**Analysis**

The approach to this project was essentially to take the perspective of an aspiring film creator looking to give themselves the best chance to make a successful film, having both high profits and good reviews. This analysis could be useful for anyone who has an eye for film but not an understanding of the specifics about the financials of the movie industry.

Because different films have widely varying budgets the financials will be split into both high revenue films, and high return on investment films. To begin the focal point would be on high ROI films. An ROI column was created by subtracting budget from revenue and dividing the difference by budget. Keep in mind that these data are from 2017 and may be a bit outdated at the time of completing this report, however the principles understood from this analysis should remain relevant. The top 200 ROI films were put into their own dataframe and amongst the top films a few names may seem familiar. Particularly *Paranormal Activity* which was the single highest ROI film in this dataset at an ROI over 12,000. Other names included: *Supersize Me, Facing the Giants, Napoleon Dynamite*. Many will recognize these films as having a smaller budget but an almost cult-like following. The genres for these top 200 ROI films were compiled and can be seen below, next to the top 10 highest ROI films.

Table

Description automatically generated Chart, bar chart, histogram

Description automatically generated

Note that the top genres chosen were drama, comedy, thriller, and horror. It should be noted the massive influence Paranormal Activity had on the horror movie industry, as the camcorder style of the film was later utilized in other horror films. Because the genre category was used as insight multiple times throughout this report, I will briefly describe the processing of this column. Within the metadata csv file, the genre column was a string, of a list, of dictionaries. Each dictionary was set up as follows: {[genre\_id: id\_number, name: genre}]. So, each genre that a film could have would be a 2 key long dictionary, and most films contained multiple genres. To figure out how often these genres occurred, the dictionaries were read into a single long list. Next, this single long list underwent mapping from the itemgetter package. This would return all values, as a string, for values that had had key values equal to name. Next, using the Counter package a command called counter would return a dictionary with each string as a key and the number of occurrences of that string as the value. Finally, this dictionary could be sorted and plotted using any graphing package in python, I opted for matplotlib.

Now the perspective will slightly shift, as we investigate what make a huge revenue film. ROI will no longer be looked at and revenue will be the focal point. In the movie industry, making over a billion dollars is surprisingly commonplace, so I shortened the dataframe to only include films that had revenues in excess of $500 million. Note: only films after 2000 are included. Only a couple attributes were found to be consistent, however one is very telling. All films that made over $500 million were roughly between 90 and 180 minutes. It could be speculated that people do not want to pay money for a short film and additionally they don’t want to sit for a 4-hour movie. This seems logical. The next common attribute of these films may shatter the dreams of some aspiring directors. But the most common attribute of these high revenue films is that they are produced by Disney/another large conglomerate. A pie-chart will be included that shows the production companies with the greatest number of high-revenue films. Many of these are Disney/owned by Disney and it appears that is an almost required pre-requisite to have such large revenue numbers for a film. There are of course outliers, but the general trend is that if it had revenue over half a billion dollars it’s likely that Disney made the film.

Chart, scatter chart

Description automatically generated Chart, pie chart

Description automatically generated

Finally, it was inferred that a larger budget should in general equate to a larger revenue. However, plotting this gave insight and opened up further investigation. The scatterplot of budget vs. revenue can be seen below (Note: the y=x line isn’t a trendline, instead it is the break-even point. All films above the line made money, all films below the line lost money).

Chart, scatter chart

Description automatically generated

A few conclusions can be drawn from this revenue vs. budget scatterplot. For one, there is a general trend that the higher the budget for the film, the more revenue will be generated. Not always true, but there is a general trend. Next if we look at the break-even line, we can see that most films (at least in this dataset) actually end up making money (69.14%). This observation inspired a look into what the less successful films had in common, so we know what to avoid.

The less successful films were grouped together using the net revenue column. The revenue was subtracted from the budget and any negative values were placed into their own dataframe. From here common attributes of the less successful films could be investigated. Most telling about these films was the country that they were produced in. I subtracted the count of films that had a positive net revenue from the count of films that had a negative net revenue, grouped by production country. Which yielded the following 2 charts.

Table

Description automatically generated Table

Description automatically generated

Note that Profitable Count is the number of positive net revenue films minus the number of negative net revenue films per country. This means that Luxembourg had 9 more films that lost money than that made money. While the United Kingdom had 144 more films that made money than that lost money. This should indicate to our aspiring film maker that countries such as Luxembourg, Belgium, Switzerland may not be the best location to film/release your movie into. Better options would be countries such as the U.S., United Kingdom, and India. This makes intuitive sense as film is a large part of all of those country’s cultures. Next the genre column was looked at again and can be seen below.

We can see, much like the high ROI films, that drama, comedy, and thriller are highly picked films. It should be noted however that drama is more highly represented in this bar chart than in the high ROI films. Also very importantly notice that horror films are very underrepresented here compared to the high ROI films. Suggesting that drama may be a bit of a gamble while horror might be a safer genre choice for your movie.

Chart, bar chart, histogram

Description automatically generated

Finally, analysis of positively reviewed films will be looked at. For this, a dataframe containing only movies that had been reviewed at least 50 times was made and a few insights were produced. This began by investigating the relationship between budget and rating. The scatterplot can be seen below.

Chart, scatter chart

Description automatically generated

Plotted here is a budget vs. rating graph. A few insights can be pulled from this. Firstly, if you’re an aspiring director without a large budget, there is a very large number of films with a smaller budget that have high ratings. The most significant trend that can be seen is that the higher your budget, the less likely you are to receive bad ratings. However, at a lower budget you can receive high or low ratings.

Next, the revenue vs. rating was looked at. It makes intuitive sense that if a movie does well commercially that’s indicative that it will do well with ratings. The results of this are very interesting. I swapped the x and y axis to get a better idea about the spread of ratings, and we can see that the ratings begin to resemble a normal distribution with a mean around 6.0-6.5. Even more interesting, because of this normal distribution it appears that revenue doesn’t have a visually noticeable effect on the ratings.

Chart, scatter chart

Description automatically generated

**Conclusion**

In conclusion quite a few valuable insights can be drawn from this project. When it comes to making money from your film, you should focus on high ROI instead of on a large revenue. The large revenue films are almost exclusively from massive companies such as Disney. When picking the genre for your film, horror has a high representation in high ROI films and a low representation in films that lost money. Based off this analysis, horror would be the best genre. Finally, producing your films in locations like Luxembourg, Belgium, and Switzerland are at a much higher risk of losing money, and instead the director should opt to produce their film in the U.S., United Kingdom, or India.

If the focus of the director is simply on getting high reviews, then they should rest assured that their budget won’t have a meaningful impact on their reviews. It was found that when looking at budget, the higher a budget a film had, the less likely it was to be rated poorly. However, there was a very high number of films with low budgets that were reviewed both positively and negatively. Furthermore, when looking at the revenue generated, the ratings resembled a normal distribution. This would indicate that ratings aren’t massively affected by revenue. Our aspiring director should take heed in the fact a low budget film can and often is reviewed very positively.

**DataSet Used**

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>

A copy of the dataset will be turned in with this report